Chronopolis – MetaArchive

Uniting Two Systems to Improve Inter-institutional Preservation
Chronopolis Basic Facts

• Three node federated data grid at UCSD/SDSC, NCAR and UMIACS with capacity for up to 50 TB of data per node (150 TB total)

• Using the Storage Resource Broker (SRB) for data management (moving to iRODS)

• Using BagIt file packaging format and SRB tools to transfer data

• Using Auditing Control Environment (ACE) for integrity checking
Current Chronopolis collections

Data Providers:

- **Inter-university Consortium of Political and Social Research** – preservation copy of all collections including 40 years of social science data and Census
- **California Digital Library** – political and government web crawls, Web-at-risk collection
- **SIO Explorer** – data from 50 years of research voyages
- **NCSU Libraries** -- State and local geospatial data

[Map of Chronopolis DataGrid and Data Providers]

http://chronopolis.sdsc.edu
MetaArchive Basic Facts

• The MetaArchive Cooperative is an independent, unincorporated, international membership association, established in 2004

• The Cooperative’s purpose is to support, promote, and extend the MetaArchive approach to distributed digital preservation practices (http://www.metaarchive.org)

• ...Is building Trusted Digital Repositories using the open source LOCKSS software to provide long-term care for digital materials

• Members to date are: Emory, Georgia Tech, Florida State U., Univ. Of Louisville, Virginia Tech, Auburn U., Boston Coll., Folger Shakespeare Library, Clemson U., Pontifícia Universidade Católica do Rio de Janeiro, Rice Univ., Univ. of Hull.

• Newest members: Penn State, Univ. Of South Carolina
MetaArchive Collections

• **The Southern Digital Culture Network** was established in coordination with the Library of Congress in 2004.

• **The Electronic Theses and Dissertations Network** preserves the electronic theses and dissertations of its member institutions.

• **The Transatlantic Slave Trade Network** preserves an international collection of materials relating to slave trade voyages across the Atlantic Ocean.
Our Grand Plans

• Chronopolis and MetaArchive realize the value in looking at inter-institutional preservation

• Have been pursuing informally

• Looking at ways of formalizing this process for long-term preservation goals
NSF INTEROP PROPOSAL *(submitted)*: IMPROVING INTER-INSTITUTIONAL PRESERVATION

- Project to develop tools and methods to automate exchange of data between MetaArchive Cooperative (LOCKSS-based) and Chronopolis (iRODS-based)

- Project to examine data transfer tools/protocols from:
  - California Digital Library microservices
  - iRODS protocols for data transfer
  - LOCKSS “plug-in” approach for data transfer

- Goal: A highly robust, easy to use preservation “system,” allowing digital objects to be shared between several major preservation networks in the U.S.
Focus Issues

• What does it mean to unite systems?

• Ability to export data between systems
  – Verify appropriate fixity
  – Transparency for system administrators

• Ability to track collections between systems
  – Verify collections are retrievable
  – Verify collections retain original characteristics
Technical Issues

• What are the best ways to have an SRB/iRODS datagrid and a LOCKSS PLN interact?

• What does it mean to have an active system (MetaArchive) and an archival system (Chronopolis) work together?

• What are the appropriate transfer technologies?
  – iRODS and LOCKSS native tools
  – CDL Micro-services, e.g. BagIt
BagIt

BagIt is a hierarchical file packaging format for the exchange of generalized digital content

- There is no software to install
- Consists of base directory with manifest file & subdirectory with content
- Manifest file has a row for each content file with:
  - Full path in content directory
  - A checksum for file

Holey Bags

- Have additional ‘fetch.txt’ file in base directory & empty content directory
- URLs for each content file are listed in fetch.txt file.
- Can reduce transfer time by fetching content in parallel

Procedural Issues

• What exactly are the inter-institutional ties?
  – “Just” backup?
  – Added service for our customers?
  – Will all customers want this?

• Legal issues with data owners

• MetaArchive and Chronopolis have very different management approaches. How do cross-institutional decisions get made?
The Big Win

• Important data preservation demonstration
  – No single system can solve all problems
  – No single system appeals to all user needs

• Practical, useful process for our organizations
  – Makes us individually stronger
  – Helps with things like TRAC
Contacts

• MetaArchive: http://www.metaarchive.org/

• Chronopolis: http://chronopolis.sdsc.edu/

• Tyler Walters: tyler@gatech.edu

• David Minor: minor@sdsc.edu